Mon. Not. R. Astron. Soc. **000**, 1–9 (2015) Printed 4 August 2015 (MN LATEX style file v2.2)

Estimating stellar atmospheric parameters based on LASSO and support-vector regression

Yu Lu,Xiangru Li⋆
School of Mathematical Sciences, South China Normal University, Guangzhou 510631, China;

4 August 2015

ABSTRACT

A scheme for estimating atmospheric parameters $T_{\texttt{eff}}$, log g, and [Fe/H] is proposed on the basis of Least Absolute Shrinkage and Selection Operator (LASSO) algorithm and Haar wavelet. The proposed scheme consists of three processes. A spectrum is decomposed using the Haar wavelet transform and low-frequency components at the fourth level are considered as candidate features. Then, spectral features from the candidate features are detected using the LASSO algorithm to estimate the atmospheric parameters. Finally, atmospheric parameters are estimated from the extracted spectral features using the support-vector regression (SVR) method. The proposed scheme was evaluated using three sets of stellar spectra respectively from Sloan Digital Sky Survey (SDSS), Large Sky Area Multi-object Fiber Spectroscopic Telescope (LAMOST), and Kurucz's model, respectively. The mean absolute errors are as follows: for 40 000 SDSS spectra, 0.0062 dex for log $T_{\texttt{eff}}$ (85.83 K for $T_{\texttt{eff}}$), 0.2035 dex for log g and 0.1512 dex for [Fe/H]; for 23963 LAMOST spectra, 0.0074 dex for log $T_{\texttt{eff}}$ (95.37 K for $T_{\texttt{eff}}$), 0.1528 dex for log g, and 0.1146 dex for [Fe/H]; and for 10469 synthetic spectra, 0.0010 dex for log $T_{\texttt{eff}}$(14.42K for $T_{\texttt{eff}}$), 0.0123 dex for log g, and 0.0125 dex for [Fe/H].

Key words: methods: statistical-techniques: spectroscopic-stars: atmospheres-stars:fundamental parameters

1 INTRODUCTION

The implementation of large-scale and deep sky survey programs, such as the Sloan Digital Sky Survey (SDSS: York et al. 2000; Ahn et al. 2012), Gaia-ESO Survey (GES: Gilmore et al. 2012; Randich et al. 2013) and the Large Sky Area Multi-object Fiber Spectroscopic Telescope (LAMOST/Guoshoujing Telescope; Zhao et al. 2006; Cui et al. 2012), has resulted in the collection of a large amount of stellar spectra. The bulk of this stellar spectral information necessitates utilisation of a fully automated characterisation process.

In particular, estimating the effective temperature, surface gravity, and metallicity from stellar spectra remains a fundamental problem (Wu et al. 2011; Song et al. 2012). For example, Muirhead et al. (2012) investigated the estimation of the effective temperature $T_{\texttt{eff}}$ and metallicity [M/H] for late-K and M-type planet-candidate host stars from the K-band spectra released by the Kepler Mission. In this study, a surface fitting method was used along with three spectral indices, namely the equivalent widths of NaI (2.210 μm) and CaI (2.260 μm) lines and an index describing the flux change between three 0.02μm wide bands centred at 2.245, 2.370, and 2.080 μm. Koleva et al. (2009) developed a full-spectrum fitting package, ULySS (University of Lyon Spectroscopic analysis Software), and explored its application in spectral parameterization. Manteiga et al. (2010) parameterised stellar spectra by extracting features based on Fourier analysis and wavelet decomposition as well as by constructing a mapping from feature space to parameter space using a forward neural networks (FNN) with three layers.

An automated estimation system for atmospheric parameters from a stellar spectrum is also referred to as a spectrum-parameterisation system (SPS) in related studies. An SPS system consists of two key processes, namely feature extraction and parameter estimation. Feature extraction determines representation of spectral information. Principal component analysis (PCA: Re Fiorentin et al. 2007; Bu and Pan 2015) and fast Fourier transform (FFT: Manteiga et al. 2010) are the two most frequently used feature extraction methods. Parameter estimation constructs a mapping from the extracted features of a spectrum to its atmospheric parameters. Some commonly used estimation methods are FNN (Re Fiorentin et al. 2007), Gaussian

⋆ Email: xiangru.li@gmail.com(X. Li)

process (GP: Bu and Pan 2015), support-vector regression (SVR: Li et al. 2014), etc.

This work describes a scheme for estimating atmospheric parameters from the stellar spectrum. The three processes of the proposed scheme are performed as follows. Several candidate features are initially obtained by transforming a spectrum into a low-frequency space using the Haar wavelet transform. Then, some representative candidate features are selected as spectral features using the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm (Tibshirani 1996). Finally, atmospheric parameters are estimated from extracted features using a regression method.

High-frequency components are removed in the first process. These components are usually more affected by noise than low-frequency components, which will be discussed further in section 5.3. The second process reduces the number of spectral features efficiently, a feature closely related to the estimation efficiency and exploration of representative features (Section 5.3). In the third process, the spectrum parameterization problem is investigated using four typical estimation methods: the radial basis function neural network (RBFNN: Schwenker et al. 2001), SVR (Chang & Lin 2001; Schökopf et al. 2002; Smola et al. 2004), K-nearest neighbor regression (KNNR: Altman et al. 1992) and least square regression (LSR: James et al. 2013). Experimental results indicate that SVR is superior to the other three methods.

This article is organised as follows. Section 2 describes some applied experimental data. After introducing four regression methods in Section 3, Section 4 describes a scheme for extracting spectral features. In Section 5, the proposed scheme is evaluated experimentally. Finally, Section 6 concludes this work

2 DATA

The proposed scheme was evaluated by performing experiments using three data sets: 33 963 actual spectra from LAMOST, 50 000 actual spectra from SDSS and 18 969 synthetic spectra computed from Kuruczs model. Actual spectra usually carry some disturbances from noise and preprocessing imperfections.

The proposed scheme is a statistical learning method. It basically involves the discovery of mapping from stellar spectra to their atmospheric parameters using empirical data, which is referred to as a training set in machine learning. At the same time, the performance of the mapping discovered should be evaluated objectively. Therefore, an independent set of stellar spectra is needed for this evaluation. This independent set is usually referred to as a test set. Thus, in each experiment, stellar spectra are split into two subsets, namely the training and test sets.

2.1 Actual spectra from SDSS

A set of 50 000 stellar spectra and their physical parameters collected from SDSS are selected (Abazajian et al. 2009; Yanny et al. 2009). The selected spectra span the ranges of [4088, 9740] K for $\mathrm{T_{eff}}$, [1.015, 4.998] dex for surface gravity and [-3.497, 0.268] dex for [Fe/H]. All stellar spectra are shifted to their rest frames (zero radial velocity) using the radial velocity provided by the SDSS/the Sloan Extension for Galactic Understanding and Exploration (SEGUE) Spectroscopic Parameter Pipeline (SSPP: Beers et al. 2006; Lee et al. 2008a,b, 2011; Allende Prieto et al. 2008; Smolinski et al. 2011) and rebinned to a maximal common log(wavelength) range [3.581862, 3.963961] with a sampling step of 0.0001. The common wavelength range is approximately [3818.23, 9203.67]Å. The SDSS training and test sets are labeled as S_{tr}^{SD} and S_{te}^{SD}, respectively. Without special instructions, the sizes of S_{tr}^{SD} and S_{te}^{SD} are 10 000 and 40 000, respectively.

2.2 Actual spectra from LAMOST

A set of 33 963 LAMOST stellar spectra (Luo et al. 2012) and their physical parameters from LAMOST pipeline are selected based on two signal-to-noise ratio (SNR) constraints: namely $\mathrm{SNR}_g \geqslant 20$ and $\mathrm{SNR}_r \geqslant 20$ in $g-$band and $r-$band (Luo et al. 2012) [1]. This constraint is proposed by considering the current quality stability of the LAMOST spectra. The LAMOST spectra are also shifted to their rest frames using the radial velocity provided by LAMOST SSPP and rebinned to their maximal common log(wavelength) range [3.5845, 3.9567] with a sampling step of 0.0001. The common wavelength range is approximately [3841.49, 9051.07]Å. All LAMOST spectra span the ranges [3853.2, 9927] K for $\mathrm{T_{eff}}$, [0.8920, 4.9959]dex for log g and [-2.3280, 0.9360] dex for [Fe/H]. The LAMOST training and test sets are labeled as S_{tr}^{LA} and S_{te}^{LA}, respectively. S_{tr}^{LA} consists of 10,000 stellar spectra, while S_{te}^{LA} consists of 23,963 spectra.

2.3 Synthetic spectra

A set of 18 969 synthetic spectra are calculated from Kurucz's NEWODF (new opacity distribution function) models (Castelli & Kurucz 2003) using the SPECTRUM (v2.76) package (Gray et al. 1994) as well as 830 828 atomic and molecular lines (contained in two files, luke.lst and luke.nir.lst). These atomic and molecular data are stored in the file stdatom.dat, which includes solar atomic abundances from Grevesse et al. (1998). The SPECTRUM package and the three data files can be downloaded from website [2].

The grids of the synthetic stellar spectra span parameter ranges [4000,9750] K for T_{eff} (45 values, stepsizes of 100 K between 4000 K and 7500 K and 250 K between 7750 K and 9750 K), [1, 5] dex for log g (17 values, stepsize of 0.25 dex) and [-3.6, 0.3] dex for [Fe/H] (27 values, stepsizes of 0.2 dex between -3.6 dex and -1 dex and 0.1 dex between -1 dex and 0.3 dex). The synthetic stellar spectra are also split into two subsets: a training set S_{tr}^{SY} and a test set S_{te}^{SY} consisting of 10 469 and 8 500 spectra, respectively.

[1] This constraint was not used on SDSS data.
[2] http://www.appstate.edu/ grayro/spectrum/spectrum.html

3 ESTIMATION MODELS AND EVALUATION METHODS

Here, $\boldsymbol{x}$ represents a description of a stellar spectrum, while y is the atmospheric parameter $T_{\mathtt{eff}}$, log g, or [Fe/H] of $\boldsymbol{x}$.

3.1 Estimation models

The spectrum parameterization problem is to recover the mapping f:

$$y = f(\boldsymbol{x}) \tag{1}$$

from a set of empirical data (training set).

In this work, f is estimated using four typical regression methods: RBFNN (Schwenker et al. 2001), SVR (Chang & Lin 2001; Schökopf et al. 2002; Smola et al. 2004), KNNR (Altman et al. 1992) and LSR (James et al. 2013). RBFNN and KNNR are non-linear regression methods, while LSR is a linear regression method. SVR can be implemented with a Gaussian kernel and a linear kernel. These two implementation cases are labelled as SVR_G and SVR_l, respectively. Therefore, SVR_G is a nonlinear method whereas SVR_l is a linear method.

3.2 Evaluation criteria

Suppose that $\hat{f}$ is an estimate of the spectrum parameterization mapping f and $S = \{(\boldsymbol{x}, y)\}$ is a data set, where $\boldsymbol{x}$ is a representation of a stellar spectrum and y is an atmospheric parameter of the corresponding star. The data set S can be a training or test set.

For convenient comparison with related reports, this study evaluates the performance of an estimate $\hat{f}$ using three methods: mean error (ME), mean of absolute error (MAE) and standard deviation (SD). These three evaluation methods are widely used in related studies (Re Fiorentin et al. 2007; Jofre et al. 2010; Tan et al. 2013).

Using the data set S, the three evaluation criteria can be defined as follows:

$$\mathtt{ME} = \frac{1}{n} \sum_{(\boldsymbol{x},y)\in S} (y - \hat{f}(\boldsymbol{x})), \tag{2}$$

$$\mathtt{MAE} = \frac{1}{n} \sum_{(\boldsymbol{x},y)\in S} |y - \hat{f}(\boldsymbol{x})|, \tag{3}$$

$$\mathtt{SD} = \sqrt{\frac{1}{n} \sum_{(\boldsymbol{x},y)\in S} (y - \hat{f}(\boldsymbol{x}))^2}, \tag{4}$$

where n is the amount of elements in S.

4 FEATURE EXTRACTION

4.1 Extract Candidate Features

Atmospheric parameters $T_{\mathtt{eff}}$, log g, and [Fe/H] show an evident non-linear dependence on stellar spectra (Table 6, Table 10, Table 11 in Li et al. 2014). Therefore, the spectrum parameterization problem is investigated by non-linearly transforming a spectrum before estimating atmospheric parameters. In this work, stellar spectra are transformed using a Haar wavelet (Mallat 2009) and decomposed into a series of components with different wavelengths and frequencies (time-frequency localization).

High-frequency components are usually more affected by noise than low-frequency components. Thus, this work obtains candidate features by removing high-frequency components. This process will be discussed further in Section 5.3.

In addition to the Haar wavelet, multiple alternatives for a non-linear transformation (Mallat 2009; Daubechies 1992), such as the Coiflets, Daubechies, Symmlet and biorthogonal wavelets, are available based on wavelet transform.

4.2 Refining the candidate features

Experiments indicate the presence of many redundancies in the extracted candidate features (Section 5.3). Therefore, this work proposes a scheme for detecting spectral features from the extracted candidate features using the LASSO algorithm (Tibshirani 1996).

Let $S_{tr} = \{(\boldsymbol{x}, y)\}$ be a training set (Section 2), where $\boldsymbol{x} = (x_1, \cdots, x_m)$ represents a stellar spectrum based on its candidate features, y is an atmospheric parameter of the corresponding star and m is a positive integer. The LASSO algorithm selects features using the following model:

$$\hat{\boldsymbol{w}} = arg\min_{\boldsymbol{w}} \{ \sum_{(\boldsymbol{x},y)\in S} (y - \sum_{i=1}^{m} w_i x_i)^2 + \lambda \sum_{i=1}^{m} \|w_i\| \}, \tag{5}$$

where $\lambda > 0$ is a preset parameter. In this model, only a few $\hat{w}_i$ will be non-zero, and the variables x_i with a non-zero $\hat{w}_i$ are selected as spectral features. The parameter λ controls the amount of non-zero parameters $\hat{w}_i$ or, equivalently, the number of detected features. In this work, the parameter λ is estimated by tenfold cross validation (Tibshirani 1996; Sjöstrand et al. 2005).

Based on the training set from SDSS (Section 2.1), 17 spectral features are detected for $\mathrm{T}_{\mathtt{eff}}$, 24 spectral features for log g, and 25 features for [Fe/H] (Table 1).

5 EXPERIMENTS AND DISCUSSION

5.1 Performance for SDSS spectra

From the detected features in Table 1, a spectrum parameterization model can be learned from the training set S_{tr}^{SD} (Section 2.1). The performance obtained using the test set S_{te}^{SD} is presented in Table 2.

For the SDSS test set, MAE errors are 0.0062, 0.2035 and 0.1512 dex for log $T_{\mathtt{eff}}$, log g and [Fe/H], respectively. To compare the proposed scheme with those in previous related reports, the performance of the proposed scheme was also evaluated using measures ME and SD. More experimental results are presented in Table 2 as well as Figs. 1 and 2. Direct comparisons with reports are shown in Section 6 and more discussion on the dispersion in Fig. 1 is presented in Section 5.4.

Table 1. Extracted features for estimating atmospheric parameters. WP is the wavelength position represented by a two-dimensional vector $[a, b]$, where a and b are the starting and ending wavelengths(Å).

(a) Extracted features for estimating $T_{\mathtt{eff}}$.

Label	WP(Å)	Label	WP(Å)	Label	WP(Å)	Label	WP(Å)
T_1	[3932.439, 3946.045]	T_2	[4217.567, 4232.159]	T_3	[4780.375, 4796.915]	T_4	[4851.343, 4868.128]
T_5	[5033.406, 5050.821]	T_6	[5070.631, 5088.175]	T_7	[5108.131, 5125.804]	T_8	[5126.984, 5144.723]
T_9	[5145.908, 5163.712]	T_{10}	[5164.901, 5182.771]	T_{11}	[5203.098, 5221.100]	T_{12}	[6562.389, 6585.094]
T_{13}	[8524.364, 8553.857]	T_{14}	[8650.914, 8680.845]	T_{15}	[8747.058, 8777.321]	T_{16}	[8844.270, 8874.870]
T_{17}	[9008.697, 9039.866]						

(b) Extracted features for estimating log g.

Label	WP(Å)	Label	WP(Å)	Label	WP(Å)	Label	WP(Å)
L_1	[3818.229, 3831.440]	L_2	[3889.215, 3902.672]	L_3	[3932.439, 3946.045]	L_4	[4095.076, 4109.245]
L_5	[4295.978, 4310.841]	L_6	[4540.064, 4555.772]	L_7	[4556.821, 4572.587]	L_8	[4573.640, 4589.464]
L_9	[4658.671, 4674.789]	L_{10}	[4833.503, 4850.226]	L_{11}	[4851.343, 4868.128]	L_{12}	[4869.249, 4886.096]
L_{13}	[4887.221, 4904.130]	L_{14}	[4923.365, 4940.399]	L_{15}	[5164.901, 5182.771]	L_{16}	[5183.964, 5201.900]
L_{17}	[5222.302, 5240.371]	L_{18}	[5241.577, 5259.712]	L_{19}	[5280.341, 5298.611]	L_{20}	[5299.831, 5318.167]
L_{21}	[5319.392, 5337.796]	L_{22}	[5418.287, 5437.033]	L_{23}	[5498.725, 5517.750]	L_{24}	[6562.389, 6585.094]

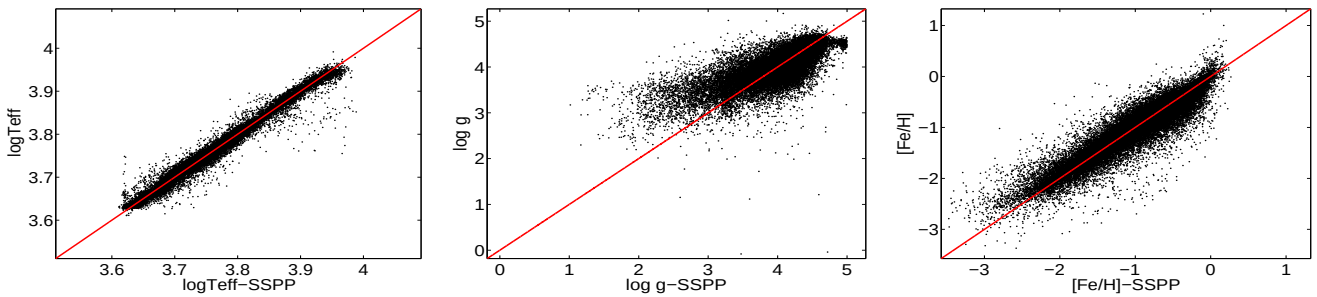
(c) Extracted features for estimating [Fe/H].

Label	WP(Å)	Label	WP(Å)	Label	WP(Å)	Label	WP(Å)
F_1	[3932.439, 3946.045]	F_2	[3990.819, 4004.626]	F_3	[4005.549, 4019.407]	F_4	[4020.333, 4034.242]
F_5	[4035.172, 4049.133]	F_6	[4506.735, 4522.327]	F_7	[4607.465, 4623.406]	F_8	[4745.282, 4761.700]
F_9	[4780.375, 4796.915]	F_{10}	[4798.019, 4814.620]	F_{11}	[4815.729, 4832.390]	F_{12}	[4851.343, 4868.128]
F_{13}	[4869.249, 4886.096]	F_{14}	[4941.536, 4958.633]	F_{15}	[4959.775, 4976.935]	F_{16}	[5051.984, 5069.463]
F_{17}	[5108.131, 5125.804]	F_{18}	[5241.577, 5259.712]	F_{19}	[5260.924, 5279.126]	F_{20}	[5280.341, 5298.611]
F_{21}	[5299.831, 5318.167]	F_{22}	[5398.362, 5417.039]	F_{23}	[5438.285, 5457.101]	F_{24}	[8524.364, 8553.857]
F_{25}	[8650.914, 8680.845]						

Table 2. Performance of the proposed scheme on 40 000 test spectra from SDSS (10 000 SDSS spectra for training, Section 2.1)

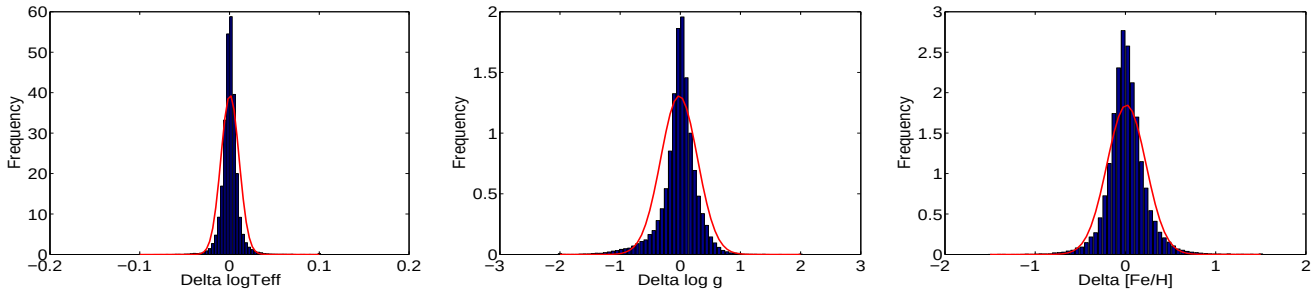
Method		log $\mathtt{Teff}$($\mathtt{Teff}$)			log g			[Fe/H]	
Parameter	MAE	ME	SD	MAE	ME	SD	MAE	ME	SD
RBFNN	0.0065(88.48)	4.42×10^{-4}(6.28)	0.0107(148.04)	0.2159	0.0205	0.3228	0.1547	6.04×10^{-4}	0.2197
SVR_G	0.0062(85.83)	6.05×10^{-4}(9.40)	0.0101(146.66)	0.2035	-0.0193	0.3053	0.1512	1.19×10^{-2}	0.2158
KNNR	0.0069(94.77)	-8.39×10^{-4}(-10.13)	0.0109(154.62)	0.2178	-0.0370	0.3069	0.2198	-3.56×10^{-2}	0.2999
LSR	0.0072(99.22)	3.45×10^{-4}(5.46)	0.0111(160.73)	0.2594	0.0270	0.3574	0.1786	3.61×10^{-3}	0.2472
SVR_l	0.0070(96.77)	3.41×10^{-4}(7.11)	0.0111(162.79)	0.2417	0.0475	0.3648	0.1758	-8.62×10^{-3}	0.2466

Notes. The unit for $T_{\mathtt{eff}}$ is K; The unit for log$T_{\mathtt{eff}}$ is log(K).



(a) log $T_{\mathtt{eff}}$ (b) log g (c) [Fe/H]

Figure 1. Performance of the proposed scheme on 40 000 test spectra from SDSS (10 000 SDSS spectra for training, Section 2.1) using SVR_G



(a) log T_{eff} (b) log g (c) [Fe/H]

Figure 2. Residual distributions of the proposed scheme on 40 000 test spectra from SDSS (10 000 SDSS spectra for training, Section 2.1)

5.2 Performance for LAMOST spectra and synthetic spectra

The proposed scheme is also tested on actual spectra from LAMOST (section 2.2) and synthetic spectra (Section 2.3).

The performances of the scheme on LAMOST spectra is presented in Table 3, as well as Figs. 3 and 4. The test results using synthetic spectra are presented in Table 4, as well as Figs. 5 and 6. The results in Tables 2, 3 and 4 show that the SVR_G and RBFNN are more suitable than KNNR, LSR and SVR_l for estimating atmospheric parameters.

5.3 Filtering and Selection - Positive or Negative?

The proposed scheme extracts spectral features by removing high-frequency components from the Haar wavelet transform and rejecting most low-frequency components by the LASSO algorithm. In this study, we determine whether these processes eliminate important spectral information, such as weak lines.

Four experiments are conducted and the results are listed in Table 5. The results show the possibility of eliminating useful spectral information. However, in application, the observed spectrum is inevitably contaminated with noise. In theory, weak lines should be more sensitive to noise.

Therefore, the loss of the elimination is trivial. The wavelet components with the lowest frequency are traditional choices for spectral features for estimating atmospheric parameters (Lu et al. 2013). In the experiments for T_{eff}, when all low-frequency wavelet components are used, the number of features will increase from 17 to 239 (increase (239-17)/17 = 1305.88 percent), but the MAE can only decrease by 0.0007 dex (11.29 percent: Experiment 1 and 2). When no component is eliminated while estimating T_{eff}, the number of features will increase from 17 to 3823 (an increase of 22388.23 percent), but the MAE error increases by 0.0398 dex (641.93 percent). A small number of detected features indicates an efficient process for estimating atmospheric parameters from spectral features. The above results suggest that the model developed in this work can estimate stellar atmospheric parameters with high accuracy.

5.4 Knowledge base, dispersion and performance

The proposed scheme is a statistical learning method. Its primary principle is to discover automatically the mapping from a stellar spectrum to its atmospheric parameters from a training set. The training set is the carrier of knowledge and affects the accuracy of the scheme.

Therefore, the size of the training set affects the performance of the proposed scheme. For example, if the size of the training set is increased from 10 000 to 15 000, 20 000, 25 000 in the experiments on SDSS spectra[3], the test dispersion can be clearly improved (Fig. 7). Similar experiments are conducted on synthetic spectra and the corresponding results are presented in Fig. 8.

Actuall data usually present some disturbances arising from noise and pre-processing imperfections (e.g., sky lines and/or cosmic ray removal residuals, residual calibration defects and interstellar extinction instability [4]). The negative effect from these factors can be reduced to a certain extent by enriching the knowledge carrier, i.e. the training set (Figs. 7 and 8).

5.5 Compactness

For simplicity, this section considers the features of log g as an example to discuss compactness. Other features of T_{eff} and [Fe/H] can be discussed similarly.

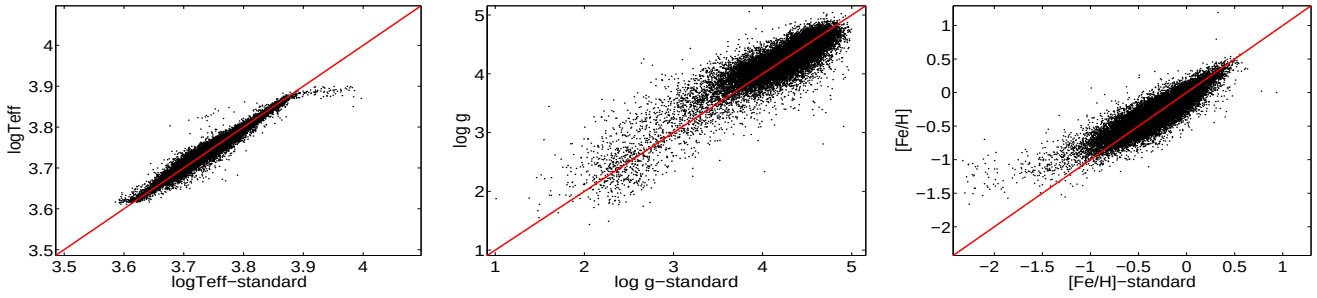
The original SDSS spectra are described by 3,821 fluxes. To estimate log g, 24 features are detected and data reduction is (3821 - 24)/(3821)≈99.37 percent. This result indicates that log g can be estimated from a spectral description with a dimension of 24 instead of 3821. The small number of features also implies high efficiency in estimating the parameter from spectral information.

Compactness indicates the study of whether the number of features can be reduced further. Experimental results show that if seven features, namely L1, L8, L13, L19, L20, L21 and L23, are rejected, the feature number decreases by 7/24 = 29 percent and the MAE error increases by 0.0027 dex (approximately 0.0027/0.2035 ≈ 1.32 percent).[5] Therefore, the feature number can be refined further if a slight decrease in accuracy is accepted.

[3] Correspondingly, the size of the test set decreases from 40 000 to 35 000, 30 000, and 25 000, respectively.

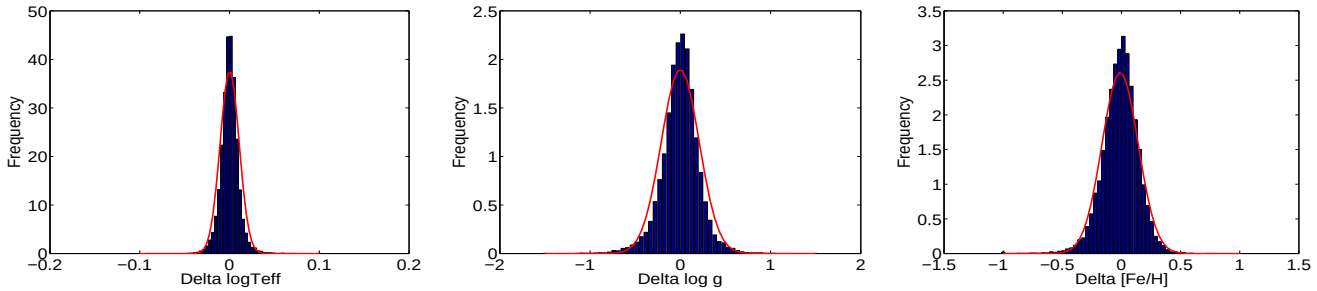
[4] By instability, we mean that a slight difference in the interstellar extinction of multiple stars may be observed.

[5] These experiments are conducted on SDSS spectra (Section 2.1).



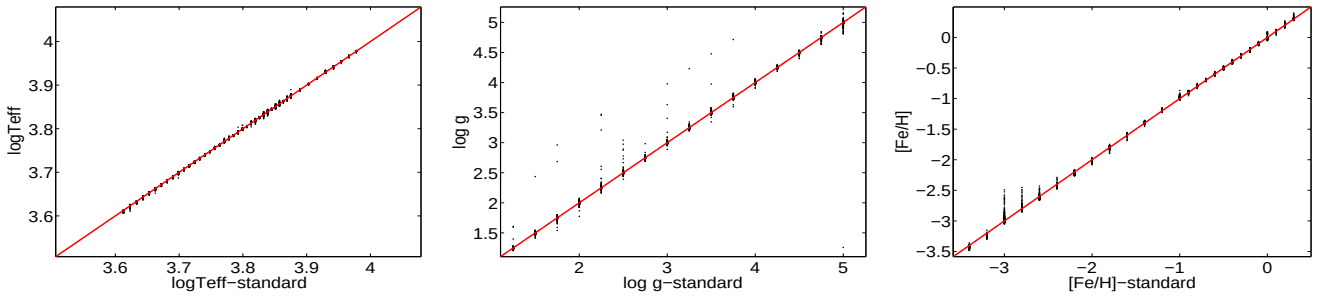
(a) log T_{eff} (b) log g (c) [Fe/H]

Figure 3. Performance of the proposed scheme on 23,963 test spectra from LAMOST (10,000 LAMOST spectra for training, Section 2.2) using SVR_G



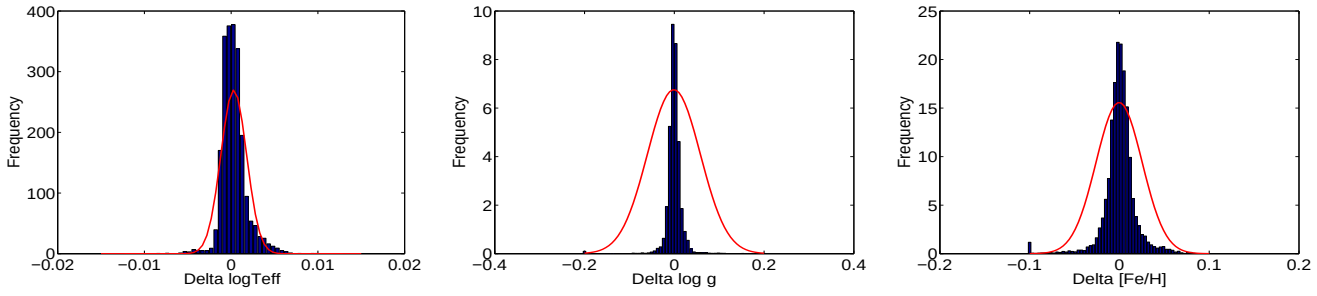
(a) log T_{eff} (b) log g (c) [Fe/H]

Figure 4. Residual distributions of the proposed scheme on 23,963 test spectra from LAMOST (10000 LAMOST spectra for training, Section 2.2) using SVR_G



(a) log T_{eff} (b) log g (c) [Fe/H]

Figure 5. Performance of the proposed scheme on 10,469 test spectra computed from KURUCZ's model (8,500 synthetic spectra for training, Section 2.3) using SVR_G



(a) log T_{eff} (b) log g (c) [Fe/H]

Figure 6. Residual distributions of the proposed scheme on 10,469 test spectra computed from KURUCZ's model (8,500 synthetic spectra for training, Section 2.3) using SVR_G

Table 3. Performance of the proposed scheme on 23 963 test spectra from LAMOST (10 000 LAMOST spectra for training, Section 2.2)

Method		log T$_{\mathtt{eff}}$(T$_{\mathtt{eff}}$)			log g			[Fe/H]	
Parameter	MAE	ME	SD	MAE	ME	SD	MAE	ME	SD
RBFNN	0.0070(91.14)	6.75×10^{-5}(2.37)	0.0099(131.36)	0.1664	0.0109	0.2753	0.1197	-0.0038	0.1767
SVR$_G$	0.0074(95.37)	1.27×10^{-4}(4.30)	0.0106(141.62)	0.1528	-0.0008	0.2102	0.1146	-0.0112	0.1528
KNNR	0.0085(111.80)	4.47×10^{-4}(10.08)	0.0126(173.66)	0.1934	0.0167	0.2730	0.1625	-0.0154	0.2151
LSR	0.0082(106.51)	6.67×10^{-4}(11.46)	0.0117(161.83)	0.2218	0.0173	0.3404	0.1312	-0.0143	0.1807
SVR$_l$	0.0081(105.73)	1.89×10^{-4}(5.44)	0.0116(159.83)	0.2070	-0.0124	0.3145	0.1311	-0.0222	0.1806

Notes. The unit for $T_{\mathtt{eff}}$ is K; The unit for log$T_{\mathtt{eff}}$ is log(K).

Table 4. Performance of the proposed scheme on 10 469 test spectra computed from KURUCZ's model (8 500 synthetic spectra for training, Section 2.3)

Method		log T$_{\mathtt{eff}}$(T$_{\mathtt{eff}}$)			log g			[Fe/H]	
Parameter	MAE	ME	SD	MAE	ME	SD	MAE	ME	SD
RBFNN	0.0010 (14.15)	1.34×10^{-4}(1.42)	0.0014(20.26)	0.0217	1.89×10^{-3}	0.0582	0.0203	1.95×10^{-3}	0.0282
SVR$_G$	0.0010 (14.42)	2.63×10^{-4}(3.34)	0.0015(20.81)	0.0123	-9.42×10^{-4}	0.0590	0.0125	-3.58×10^{-4}	0.0256
KNNR	0.0027(39.39)	3.92×10^{-5}(0.19)	0.0041 (61.08)	0.2167	3.55×10^{-2}	0.3166	0.1007	2.90×10^{-2}	0.1611
LSR	0.0026 (36.18)	-2.54×10^{-4}(-3.68)	0.0033(46.05)	0.1416	2.36×10^{-2}	0.1902	0.0903	9.77×10^{-3}	0.1175
SVR$_l$	0.0025 (34.91)	1.14×10^{-4}(1.79)	0.0032 (45.80)	0.1343	2.21×10^{-2}	0.1920	0.0783	4.12×10^{-3}	0.1122

Notes. The unit for $T_{\mathtt{eff}}$ is K; The unit for log$T_{\mathtt{eff}}$ is log(K).

6 CONCLUSION AND FUTURE WORK

This work investigated estimation of effective temperature (T$_{\mathtt{eff}}$), surface gravity (log g) and metallicity ([Fe/H]) from stellar spectra based on the Haar wavelet transform and LASSO algorithm. The proposed scheme is evaluated using actual spectra from SDSS and LAMOST as well as synthetic spectra computed from Kurucz's model. Favorable results are achieved in all cases.

The proposed scheme exhibits excellent robustness and sparseness. The features are extracted using two steps. From the SDSS data, the original spectra, described by 3821 fluxes, are initially decomposed using the Haar wavelet transform and low-frequency coefficients (239 features) are selected as candidate features. In this step, some noises and redundancies with high-frequency components are removed. Then a small subset of the candidate features is chosen as spectral features using the LASSO algorithm. The second step is a supervised learning process that selects features according to their correlation with the parameter to be estimated. The number of selected features is 17 for T$_{\mathtt{eff}}$, 24 for log g and 25 for [Fe/H]. A representative work is Re Fiorentin et al. (2007), in which 50 features are extracted for estimating atmospheric parameters.

Another advantage of the proposed scheme is its high accuracy. Using the SVR$_G$ method and 40 000 stellar spectra from SDSS, the MAEs are 0.0062 dex for log T$_{\mathtt{eff}}$ (85.83 K for T$_{\mathtt{eff}}$), 0.2035 dex for log g and 0.1512 dex for [Fe/H]. Further details are shown in Table 2. In previous reports, Re Fiorentin et al. (2007) estimated the parameters of 19 000 spectra from SDSS with MAEs of 0.0126 dex for log T$_{\mathtt{eff}}$, 0.3644 dex for log g and 0.1949 dex for [Fe/H]. Jofre et al. (2010) first highly compressed the data using a likelihood method and then estimated the parameters T$_{\mathtt{eff}}$, log g and [Fe/H] from low-resolution stellar spectra measured by SEGUE; the standard deviations (SDs) of the errors were 130 K for T$_{\mathtt{eff}}$, 0.5 dex for log g and 0.25 dex for [Fe/H]. Therefore, the results estimated using the proposed scheme exhibit higher accuracy compared with those reported in literature.

In this work, the proposed scheme is evaluated using three different data sets (SDSS, LAMOST and synthetic spectra). On each kind of data, the proposed model is learned and tested independently. An interesting problem is the estimation of atmospheric parameters of one data (e.g., LAMOST) using a model learned from the other data sets(e.g., SDSS or synthetic spectra). For example, Re Fiorentin et al. (2007) investigated how to estimate the atmospheric parameters of SDSS data from synthetic spectra and vice versa. In this process, residual calibration defects should be considered. This work focuses on sparse feature extraction and the abovementioned problem will be investigated in future work.

ACKNOWLEDGMENTS

The authors would like to thank the reviewer and editor for their instructive comments and extend their thanks to Professor Ali Luo and Fang Zuo for their support and discussions. This work is supported by the National Natural Science Foundation of China (grant No: 61273248, 61075033, 61202315), the Natural Science Foundation of Guangdong Province (2014A030313425,S2011010003348), the Open Project Program of the National Laboratory of Pattern Recognition(NLPR) (201001060) and the high-performance computing platform of South China Normal University..

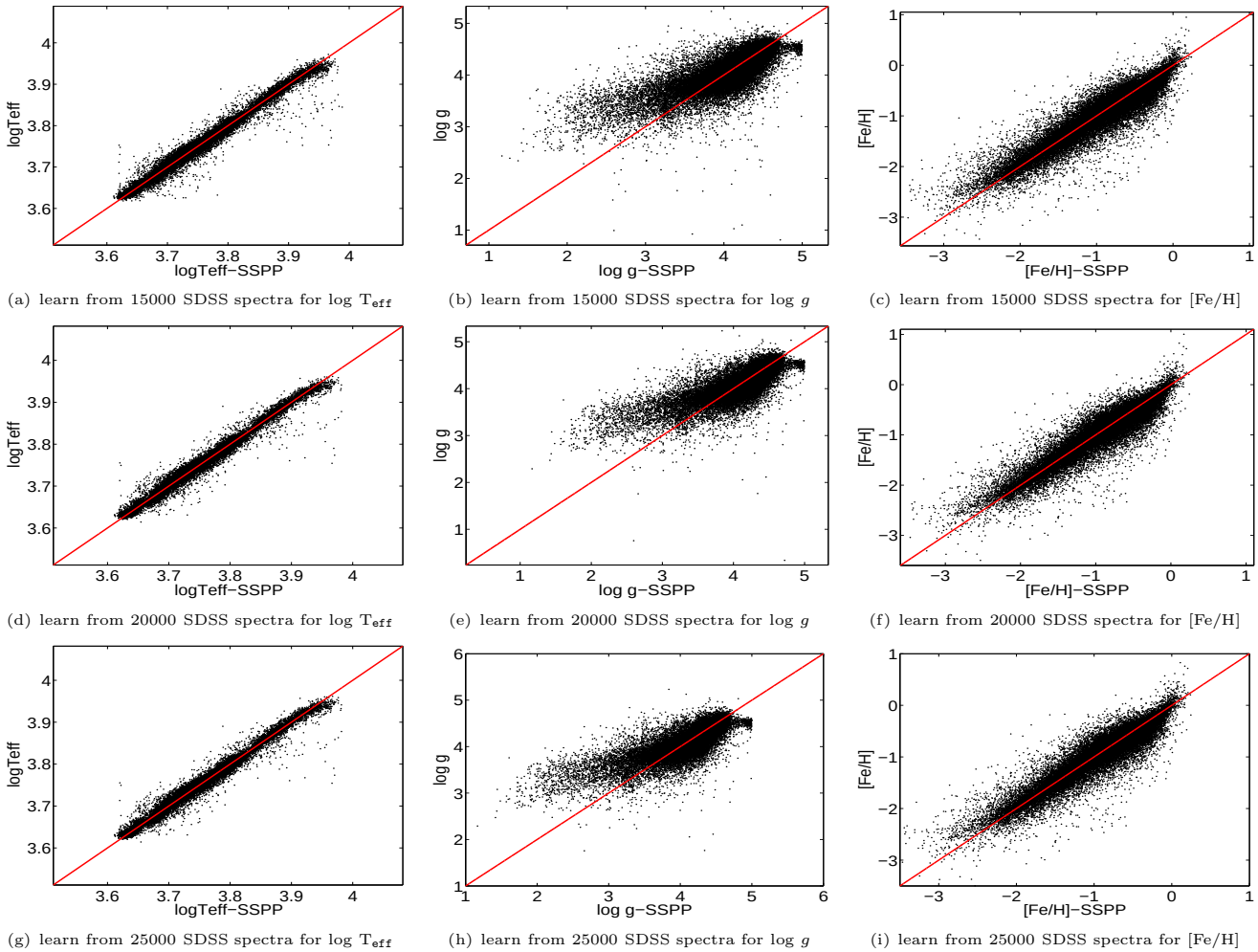
REFERENCES

Abazajian K.N., Adelman-McCarthy J.K. , Agüeros M.A., Allam S.S., Allende Prieto C., An D., Anderson K.S.J.,

Table 5. In these experiments, the advantages and disadvantages of eliminating high-frequency, as well as many low-frequency, components are evaluated. The parameters are estimated by support vector machine (RBF kernel) and RBF neural network on SDSS samples. Their performances are assessed by MAE. WT(i,0) and WT(i,1) represent the coefficients of a wavelet transform with $i-$level decomposition on the approximation and high-frequency sub-bands. $\{T_i\}$, $\{L_i\}$, and $\{F_i\}$ denote the extracted features for log $T_{\mathtt{eff}}$, log g, and [Fe/H], respectively. The number behind ":" represents the total number of the utilised features.

log $T_{\mathtt{eff}}$			log g			[Fe/H]		
Features	SVR_G	RBFNN	features	SVR	RBFNN	features	SVR_G	RBFNN
$\{T_i\}$:17	0.0062	0.0065	$\{L_i\}$:24	0.2035	0.2159	$\{F_i\}$:25	0.1512	0.1547
WT(4,0):239	0.0055	0.0062	WT(4,0):239	0.1909	0.2267	WT(4,0):239	0.1311	0.1486
WT(4,1)+$\{T_i\}$:256	0.0165	0.0083	WT(4,1)+$\{L_i\}$:263	0.2368	0.2449	WT(4,1)+$\{F_i\}$:264	0.1862	0.1770
Full:3823	0.0460	0.0131	Full:3823	0.3726	0.2366	Full:3823	0.4118	0.1769

Notes. The unit for $T_{\mathtt{eff}}$ is K; The unit for log$T_{\mathtt{eff}}$ is log(K).



(a) learn from 15000 SDSS spectra for log $T_{\mathtt{eff}}$ (b) learn from 15000 SDSS spectra for log g (c) learn from 15000 SDSS spectra for [Fe/H]

(d) learn from 20000 SDSS spectra for log $T_{\mathtt{eff}}$ (e) learn from 20000 SDSS spectra for log g (f) learn from 20000 SDSS spectra for [Fe/H]

(g) learn from 25000 SDSS spectra for log $T_{\mathtt{eff}}$ (h) learn from 25000 SDSS spectra for log g (i) learn from 25000 SDSS spectra for [Fe/H]

Figure 7. Dispersion can be improved by increasing the size of training set from 10 000 to 15 000, 20 000, 25 000 in the experiments on SDSS test spectra. This experiment is conducted using SVR_G.

Anderson S.F., et al., 2009, ApJS, 182, 543

Ahn C.P., Alexandroff R., Allende Prieto C., Anderson S.F., Anderton T., Andrews B.H., Aubourg É. , Bailey S., et al., 2012, ApJS, 203, 21

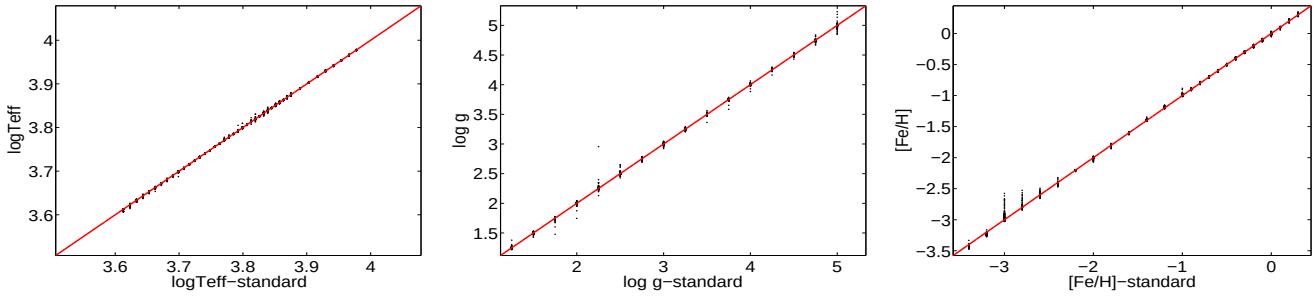
Allende Prieto C., Sivarani T., Beers T.C., Lee Y.S., Koesterke L., Shetrone M., Sneden C., Lambert D.L., et al., 2008, AJ, 136, 2070

Altman N.S. 1992, The American Statistician, 46, 175

Beers T.C., Lee Y.S., Sivarani T., Allende Prieto C., Wilhelm R., Re Fiorentin P., Bailer-Jones C., Norris J.E., et al., 2006, Mem. Soc. Astron. Ital., 77, 1171

Bu, Y., Pan, J. 2015, MNRAS, 447, 256

Castelli, F., & Kurucz, R. L. 2003, in IAU Symp. 210, Modelling of Stellar Atmospheres, ed. N. E. Piskunov, W.



(a) log $T_{\mathtt{eff}}$ (b) log g (c) [Fe/H]

Figure 8. Dispersion can be improved by increasing the size of training set from 8 500 to 10 469 in the experiment on synthetic test spectra. This experiment is conducted using SVR_G.

W. Weiss, & D. F. Gray (Cambridge: Cambridge Univ. Press), A20

Chang, C.C., Lin, C.J. 2001, LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm

Cui X., Zhao Y., Chu Y., Li G., Li Q., Zhang L., Su H., Yao Z., et al. 2012, Res. Astron. Astrophys., 12,1197

Daubechies, I. 1992, Ten Lectures on Wavelets (Philadelphia: Society for Industrial and Applied Mathematics)

Gray R.O., Corbally C.J. 1994, AJ, 107, 742

Grevesse N., Sauval A.J., 1998, Space Sci. Rev., 85, 161

Gilmore G., Randich S., Asplund M., Binney J., Bonifacio P., Drew J., Feltzing S., Ferguson A., et al. 2012, Messenger, 147, 25

James, G., Witten, D., Hastie, T., & Tibshirani, T. 2013, An Introduction to Statistical Learning with Applications in R (New York: Springer-Verlag)

Jofre P., Panter B., Hansen C.J., Weiss A., et al., 2010, A&A, 517, A57

Koleva M., Prugniel P., Bouchard A., Wu Y., 2009, A&A, 501, 1269

Lee Y.S., Beers T.C., Sivarani T., Allende Prieto C., Koesterke L., Wilhelm R., Re Fiorentin P., Bailer-Jones C.A.L., et al, 2008a, AJ, 136, 2022

Lee Y.S., Beers T.C., Sivarani T., Johnson J.A., An D., Wilhelm R., Allende Prieto C., Koesterke L., et al., 2008b, AJ, 136, 2050

Lee Y.S., Beers T.C., Allende Prieto C., Lai D.K., Rockosi C.M., Morrison H.L., Johnson J.A., An D., Sivarani T., Yanny B., 2011, AJ, 141, 90

Li X., Wu Q. M. J., Luo A., Zhao Y., Lu Y., Zuo F., Yang T., Wang Y., 2014, ApJ, 790, 105

Lu, Y., Li, X., Wang, Y., Yang, T. 2013, Spectroscopy and Spectral Analysis,, 33(7), 2010

Luo A., Zhang H., Zhao Y., Zhao G., Cui X., Li G., Chu Y., Shi J., et al., 2012, Res. Astron. Astrophys., 12, 1243

Mallat, S. 2009, A Wavelet Tour of Signal Processing (3rd ed.; Boston: Academic Press)

Manteiga M., ORDÓÑEZ D., Dafonte C., ARCAY B., 2010, PASP122, 608

Muirhead P.S., Hamren K., Schlawin E., Rojas-Ayala B., Covey K.R., Lloyd J.P., 2012, ApJL, 750, L37

Randich, S., Gilmore, G., Gaia-ESO Consortium. 2013, The Messenger, 154, 47

Re Fiorentin P., Bailer-Jones C.A.L., Lee Y.S., Beers T.C., Sivarani T., Wilhelm R., Allende Prieto C., Norris J.E., 2007, A&A, 467, 1373

Schölkopf, B., Smola, Alex J. 2002, MIT Press, Learning with Kernels

Schwenker, F., Kestler, H.A., Palm, G. 2001, Neural Networks, 14(4-5), 439

Sjöstrand, K. Matlab Implementation of LASSO, LARS, the Elastic Net and SPCA (Version 2.0). Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2005.6, http:// www2.imm.dtu.dk/pubdb/p.php?3897

Smola, Alex J., Schölkopf, B. 2004, Statistics and Computing, 14(3), 199

Smolinski J.P., Lee Y.S., Beers T.C., An D., Bickerton S.J., Johnson J.A., Loomis C.P., Rockosi C.M., Sivarani T., Yanny B., 2011, AJ, 141, 89

Song Y., Luo A., Comte G., Bai Z., Zhang J., Du W., Zhang H., Chen J., Zuo F., Zhao Y., 2012, Res. Astron. Astrophys., 12, 453

Tan X., Pan J., Wang J., Luo A., Tu L., 2013, Spectrosc. Spectral Anal., 33, 1397

Tibshirani, R. , 1996, J. R. Stat. Soc. B, 58, 267

Wu, Y. , Luo, A. , Li, H. , et al. 2011. RAA, 11, 924 Wu Y., Luo A., Li H., Shi J., Prugniel P., Liang Y., Zhao Y., Zhang J., et al., 2011, Res. Astron. Astrophys., 11, 924

Yanny B., Rockosi C., Newberg H.J., Knapp G.R., Adelman-McCarthy J.K., Alcorn B., Allam S., Allende Prieto C., et al., 2009, AJ, 137, 4377

York D.G., Adelman J., Anderson J.E.Jr., Anderson S.F., Annis J., Bahcall N.A., Bakken J.A., Barkhouser R.et al., 2000, AJ, 120, 1579

Zhao G., Chen Y., Shi J., Liang Y., Hou J., Chen L., Zhang H., Li A. 2006, Chin. J. Astron. Astrophys., 6, 265

This paper has been typeset from a TEX/ LATEX file prepared by the author.